

DataXperiment Sachbericht: CLIMB Projekt

1. Einleitung

Ziel des Projekts CLIMB war die Entwicklung eines angepassten CLIP-Modells zur Analyse von 3D-MRT-Daten im Bereich der Brustkrebsdiagnostik. Die im Rahmen des Projekts durchgeführten Arbeiten, die erzielten Ergebnisse sowie die Verwendung der bereitgestellten Mittel werden im nachfolgenden aufgelistet.

2. Durchführung der Arbeitspakete

AP1: Anpassung von CLIP für die Analyse von 3D-MRT-Daten (Monat 1-3)

Im Rahmen dieses Vorhabens wurde das CLIP-Modell für die Verarbeitung medizinischer 3D-MRT-Daten angepasst und erweitert. Der Fokus lag auf der Optimierung der Verarbeitung volumetrischer Bilddaten sowie der Integration domänenspezifischer Sprachmodelle zur Verbesserung der multimodalen Repräsentation. Die ursprüngliche CLIP-Architektur, die primär für 2D-Bilddaten entwickelt wurde, wurde entsprechend modifiziert.

Zur Verarbeitung der Bildinformationen wurde der ursprüngliche Vision Encoder durch den *Medical Slice Transformer* (MST) ersetzt. Dieser verarbeitet einzelne 2D-Slices eines MRT-Volumens, welche zuvor mithilfe von **DINOv2**, einem selbstüberwachten Foundation Model, in latente Repräsentationen überführt werden. Durch das umfassende domänenübergreifende Vortraining von DINOv2 lassen sich hochwertige visuelle Merkmale extrahieren, ohne dass ein aufwändiges domänenspezifisches Training auf medizinischen Bilddaten erforderlich ist. Dies reduziert den Datenbedarf erheblich und ermöglicht eine effektivere Feinabstimmung des Modells.

Für die adäquate semantische Verarbeitung medizinischer Fachsprache wurde der standardmäßige CLIP-Textencoder durch **German-MedBERT** ersetzt – ein spezialisiertes Sprachmodell, das auf deutschsprachige medizinische Terminologie optimiert ist. Dadurch wird eine präzisere Einbettung von Texteingaben in den semantischen Raum ermöglicht.

Zur Fusion der extrahierten Bild- und Textinformationen wurde ein Transformer-Modell implementiert. Dabei kam die Bibliothek **x-transformers** (<https://github.com/lucidrains/x-transformers>) zum Einsatz. Der entwickelte Transformer besteht aus lediglich vier Layern und umfasst rund 6 Millionen trainierbare Parameter. Diese vergleichsweise geringe Modellgröße wurde bewusst gewählt, um Overfitting insbesondere bei begrenzten Trainingsdaten zu vermeiden. Im Unterschied zur ursprünglichen CLIP-Implementierung wurde der Mechanismus **Rotary Positional Embeddings (RoPE)** integriert, da dieser die Konvergenzgeschwindigkeit von Transformern nachweislich verbessert.

Zusätzlich wurde eine weitere Erweiterung realisiert, bei der die transformerbasierten Module durch ein vortrainiertes **LLaMA 3.2-Modell** ersetzt werden können. Hierbei wurde die kleinste verfügbare Variante mit etwa 1 Milliarde Parametern verwendet, um ein Training auch auf Hardware mit begrenzter Kapazität – konkret auf NVIDIA L40 GPUs – zu ermöglichen. Diese Option erlaubt den Einsatz moderner, leistungsstarker Sprachmodelle bei gleichzeitig vertretbarem Ressourcenaufwand.

Darüber hinaus wurden zwei weitere Modellkonfigurationen implementiert:

1. **CLIP-CoCa-Erweiterung:** In Anlehnung an das CoCa-Modell (Contrastive Captioners, <https://arxiv.org/abs/2205.01917>) wurde ein zusätzlicher Transformer auf das bestehende CLIP-Modell aufgesetzt. Dieser agiert als autoregressives Sprachmodell, das auf Grundlage der Bildrepräsentation vollständige medizinische Befundtexte generiert.
2. **Inverses CoCa-Modell:** Aufbauend auf dem CoCa-Prinzip wurde ein weiterer Transformer implementiert, der das Bild basierend auf einem gegebenen Befund autoregressiv rekonstruiert. Ziel dieser Konfiguration ist die Evaluierung der semantischen Tiefe und Konsistenz der multimodalen Repräsentation, insbesondere in Hinblick auf deren Generativität.

AP2: Modellkonfiguration und Training (Monat 3-5)

Nach der erfolgreichen Implementierung lag der Fokus in den Monaten drei bis fünf auf der Konfiguration und dem Training der entwickelten Modellarchitektur. Ziel war es, verschiedene Konfigurationsansätze zu evaluieren, um die multimodale Repräsentationsfähigkeit des Systems zu maximieren und die Balance zwischen Modellkomplexität, Generalisierbarkeit und Spezialisierung zu finden.

Für das Training wurden drei wesentliche Konfigurationsansätze definiert:

1. **Konfiguration 1 – Contrastive Learning (klassisches CLIP-Modell):**
In dieser Basisvariante werden Bild- und Textrepräsentationen im gemeinsamen Embeddingraum so optimiert, dass semantisch passende Paare zueinander gezogen (positive Paare) und unpassende getrennt werden (negative Paare). Ziel ist es, eine robuste multimodale Ähnlichkeitsmetrik zu lernen.

Neben der Standardarchitektur wurde in dieser Konfiguration zusätzlich eine Variante getestet, bei der das domänenspezifische German-MedBERT durch ein vortrainiertes LLaMA 3.2-Modell (1B Parameter) ersetzt wurde. Dabei kam ausschließlich das Embedding-Modul von LLaMA zum Einsatz. Dieses ist für mehrere Sprachen optimiert und bietet auch für deutschsprachige Texte eine solide Grundlage, allerdings ohne spezielle medizinische Anpassungen.
2. **Konfiguration 2 – CoCa-Erweiterung (Contrastive Learning + Textgenerierung):**
Aufbauend auf dem Prinzip des CoCa-Modells wurde hier ein zusätzlicher Transformer integriert, der auf Grundlage der Bildrepräsentation autoregressiv vollständige medizinische Befundtexte generiert. Ziel ist es, nicht nur Ähnlichkeiten zu modellieren, sondern auch kontextreiche, strukturierte Sprachinformationen zu erzeugen.
3. **Konfiguration 3 – Vollständige Erweiterung (Contrastive Learning + Text- und Bildgenerierung):**
Diese umfassendste Konfiguration kombiniert alle zuvor genannten Komponenten. Zusätzlich zur Textvorhersage wurde ein weiteres autoregressives Modul integriert, das auf Basis des Textes eine Sequenz von Slice-Embeddings generiert, um so

bildliche Repräsentationen rekonstruieren zu können. Dieser bidirektionale Ansatz ermöglicht eine tiefere Kopplung von Text und Bild und eröffnet neue Evaluationsmöglichkeiten zur Konsistenzprüfung.

Das Training aller Modelle erfolgte auf dem internen GPU-Cluster unter Verwendung eines Systems mit 10 NVIDIA L40S-Grafikkarten. Die Trainingsdauer pro Konfiguration betrug ca. 24 Stunden. Als Optimierungsstrategie wurde ein Lernraten-Scheduler mit einem sogenannten Warm-up implementiert. Dabei begann das Training mit einer initialen Lernrate von $1e-7$, die innerhalb der ersten 1.000 Iterationen linear auf eine maximale Lernrate von $5e-4$ gesteigert wurde.

Die Batch Size wurde standardmäßig auf 32 gesetzt. Eine möglichst große Batch Size ist insbesondere im Kontext von Contrastive Learning vorteilhaft, da sie die Anzahl der negativen Paare im Training erhöht und somit die Qualität der Repräsentationen verbessert. Allerdings setzt die verfügbare Hardware hier eine Grenze: Bei einem GPU-Speicher von 48 GB pro NVIDIA L40S konnte die Basis-Konfiguration (reines CLIP mit MedBERT) vollständig auf einer einzelnen GPU ausgeführt werden. In den komplexeren Konfigurationen, insbesondere bei Einsatz des LLaMA-1B-Modells, überstieg der Speicherbedarf die Kapazität einer GPU, sodass die Batch Size auf zwei GPUs verteilt werden musste. In diesen Fällen kamen jeweils 16 Samples pro GPU zum Einsatz, um weiterhin eine effektive Gesamt-Batch-Größe von 32 aufrechtzuerhalten.

AP3: Modellevaluation (Monat 5-6)

Nach der Implementierung und dem erfolgreichen Training der verschiedenen Modellkonfigurationen lag der Schwerpunkt in den Monaten fünf bis sechs auf der systematischen Evaluation der Modellleistung. Dabei wurden zwei zentrale Aspekte untersucht: die diagnostische Genauigkeit des Modells sowie dessen Erklärbarkeit (Explainability).

Zur Bewertung der diagnostischen Genauigkeit wurde zunächst ein dediziertes Testset definiert, das vollständig unabhängig von den Trainingsdaten war. Für jede Untersuchung in diesem Testset wurden dem Modell zwei standardisierte Textformulierungen vorgegeben: „*Karzinom.*“ und „*Kein Karzinom.*“. Anschließend wurde die Ähnlichkeit zwischen der Bildrepräsentation und den beiden Texteingaben berechnet. Diese Ähnlichkeitswerte wurden als Klassifikationswahrscheinlichkeiten interpretiert. Das Modell sollte dabei eine höhere Ähnlichkeit zur zutreffenden Aussage aufweisen. Dieses Verfahren wurde nicht nur für Karzinome, sondern für ein erweitertes Set von neun pathologischen Befunden durchgeführt:

- *Fibroadenom*
- *Adenose*
- *Lymphknoten*
- *Zyste*
- *Fettgewebsnekrose*
- *Duktektasie*
- *Radiäre Narbe*
- *Duktales Karzinom in situ (DCIS)*
- *Karzinom*

Für jede Klasse wurde die Area Under the Curve (AUC) berechnet, basierend auf den durch das Modell erzeugten Similarity Scores. Die durchschnittliche AUC über alle Klassen diente als Maß für die diagnostische Gesamtperformance.

Zur Bewertung der Erklärbarkeit des Modells wurde auf eine Ablationsmethode basierend auf Attention-Maskierung zurückgegriffen. Hierbei wurde systematisch die interne Attention-Matrix des Vision-Transformers so modifiziert, dass das Modell jeweils eine einzelne Slice des Bildvolumens nicht mehr betrachten konnte. Für jede der so erzeugten Maskierungen wurde erneut die Ähnlichkeit zwischen dem Bild und den beiden Textformulierungen berechnet.

Dieser Prozess wurde iterativ durchgeführt, bis jede Slice einmal maskiert war. Aus der Veränderung der Similarity Scores zwischen den beiden Textformulierungen („*Karzinom.*“ vs. „*Kein Karzinom.*“) bei Maskierung einzelner Slices wurde abgeleitet, welche Slice einen besonders starken Einfluss auf die Modellentscheidung hatte. Die Ergebnisse wurden in Form einer Heatmap visualisiert, die die Slices mit dem stärksten Einfluss auf die Entscheidung hervorhebt. Diese Methode erlaubt eine intuitive Interpretation der Modellvorhersagen und liefert Hinweise darauf, welche Bildregionen für die Entscheidung besonders relevant waren.

3. Verwendung der Zuwendung und Nachweis der Notwendigkeit

- **Wissenschaftlicher Mitarbeiter:** Zuständig für die Modellentwicklung, Implementierung und Evaluierung. Diese Stelle wurde über Hausmittel der Arbeitsgruppe gedeckt.
- **Studentische Hilfskräfte:** Es wurden insgesamt drei studentische Hilfskräfte im Projekt zur Unterstützung bei der technischen Umsetzung, insbesondere bei der Datenverarbeitung beschäftigt.
- **Reisemittel:** Es wurden keine Mittel für Reisekosten verausgabt, da auf Grund der kurzen Laufzeit des Projektes eine Präsentation der Ergebnisse bei einer Konferenz zeitlich nicht machbar war.

4. Nutzen und Verwertbarkeit der Ergebnisse

Die im Rahmen dieses Projekts entwickelten Methoden und Modelle liefern sowohl in wissenschaftlicher Hinsicht als auch im Hinblick auf zukünftige Anwendungen wichtige Impulse. Die Verwertbarkeit der Ergebnisse lässt sich in drei zentrale Bereiche untergliedern: klinische Anwendbarkeit, wissenschaftlicher Nutzen sowie Perspektiven für die Weiterentwicklung.

Klinische Anwendbarkeit

Zum aktuellen Stand ist die direkte klinische Nutzung des Modells noch nicht möglich. Zwar konnten durch die multimodale Modellarchitektur erste diagnostisch relevante Zusammenhänge abgebildet werden, jedoch reichen die erreichten diagnostischen Kennwerte – insbesondere im Hinblick auf Sensitivität und Spezifität – noch nicht aus, um eine zuverlässige medizinische Entscheidungsunterstützung im klinischen Alltag zu gewährleisten. Auch die aus den Attention-Maps abgeleiteten Heatmaps sind derzeit noch nicht hinreichend konsistent oder interpretierbar, um als erklärbare visuelle Hilfsmittel für die Befundung eingesetzt zu werden. Eine unmittelbare Integration in diagnostische Prozesse ist daher aktuell nicht vorgesehen.

Wissenschaftlicher Nutzen

Die durchgeführten Experimente liefern wertvolle Erkenntnisse für die Forschung im Bereich multimodaler Repräsentationslernen im medizinischen Kontext. Besonders hervorzuheben ist die Beobachtung, dass eine bidirektionale Erweiterung des ursprünglichen CLIP-Ansatzes – also die Kombination von *Bild-zu-Text* und *Text-zu-Bild*-Vorhersagen – Vorteile für das initiale Modelltraining mit sich bringt. Dies deutet darauf hin, dass multimodale Konsistenz eine stärkere Repräsentationsfähigkeit fördern kann.

Der zentrale Nutzen für die wissenschaftliche Community liegt jedoch in der Bereitstellung des implementierten Frameworks. Der quelloffene Code wurde so gestaltet, dass er modular und erweiterbar ist und somit als Grundlage für weiterführende Projekte dient. Forscherinnen und Forscher können das Framework unkompliziert an eigene Datensätze und Aufgabenstellungen anpassen, wodurch die Ergebnisse des Projekts auch über das ursprüngliche Anwendungsszenario hinaus nachhaltig nutzbar werden.

Zukunftsperspektiven

Ziel ist die sukzessive Weiterentwicklung des Modells in Richtung eines umfassenden, KI-gestützten Assistenzsystems für die Brustkrebsfrüherkennung. Im Zuge dessen sind Schritte zur Anwendung und Integration in bestehende Forschungs- und Infrastruktursysteme vorgesehen, insbesondere im Rahmen des europäischen Projekts ODELIA.

Die entwickelten Methoden sollen perspektivisch in das Forschungsprojekt ODELIA (<https://odelia.ai>) eingebunden werden. ODELIA steht für *Open Consortium for Decentralized Medical Artificial Intelligence* und verfolgt das Ziel, datenschutzgerechte, föderierte KI-Technologien für medizinische Bildgebung und Befundung zu entwickeln. Das Projekt stellt einen Zusammenschluss von zwölf führenden europäischen Partnerinstitutionen dar, darunter Universitätskliniken, Forschungseinrichtungen und Technologiedienstleister.

In diesem Kontext ist das CLIMB Projekt aus zwei Gründen besonders vorteilhaft:

1. Erweiterte Trainingsdatenmenge:

Durch die Zusammenarbeit mehrerer europäischer Partner entsteht ein signifikant größerer und heterogenerer Datenpool als es einzelnen Institutionen möglich wäre. Dieser Datenpool stellt eine ideale Grundlage für das weitere Training des Modells dar.

2. Effizienz:

Ein wesentlicher Engpass in der KI-Entwicklung im medizinischen Bereich ist der hohe Aufwand für die manuelle Annotation von Trainingsdaten, der sowohl zeitlich als auch finanziell kostenintensiv ist. Die im CLIMB Projekt entwickelte Lösung nutzt dagegen ausschließlich Informationen aus bereits vorliegenden medizinischen Befunden. Dies ermöglicht es, das Modell schnell und kosteneffizient auf neue Fragestellungen anzupassen, ohne dass zusätzlich strukturierte Annotationen erforderlich sind.

In Kombination mit verteilten, datenschutzkonformen Lernverfahren, wie sie im Rahmen von ODELIA entwickelt werden, könnte ein solches Brust-CLIP-Modell einen zentralen Beitrag zur Realisierung datensparsamer, skalierbarer und interoperabler KI-Systeme leisten.

5. Fortschritt auf dem Forschungsgebiet

Das Projekt leistet einen konkreten Beitrag zur Weiterentwicklung von KI-gestützten Verfahren in der radiologischen Diagnostik – insbesondere im Bereich multimodaler Bild-Text-Repräsentationen für die Brustbildgebung. Im Laufe der Projektlaufzeit wurde der Stand der Technik kontinuierlich beobachtet, mit aktuellen Entwicklungen abgeglichen und gezielt in die eigene Arbeit integriert.

Vergleich mit aktuellen Entwicklungen im Bereich KI-gestützter radiologischer Diagnostik

In den letzten Jahren ist ein deutlicher Trend zur Integration multimodaler Modelle in die radiologische Diagnostik erkennbar. Während frühere Systeme primär auf rein visuelle Klassifikation oder Segmentierung fokussiert waren, gewinnen Ansätze, die Bild- und Textinformationen gemeinsam verarbeiten (z. B. BiomedGPT, BioViL, MedCLIP), zunehmend an Bedeutung. Diese Modelle ermöglichen eine flexiblere und kontextsensitivere Analyse medizinischer Bilddaten, insbesondere bei komplexen Pathologien. Das hier entwickelte Framework unterscheidet sich von vielen aktuellen Forschungsansätzen, indem es explizit auf die Verarbeitung volumetrischer (3D-) Bilddaten ausgelegt ist – ein Konzept, das zwar in der Literatur zunehmend Beachtung findet, jedoch bislang nur vereinzelt praktisch implementiert wurde.

Analyse neuerer Forschungsergebnisse während der Projektlaufzeit

Während der Projektlaufzeit bis März 2025 wurden keine vergleichbaren Modelle zur Integration von volumetrischen (3D-) Bilddaten veröffentlicht. Allerdings gab es zahlreiche neue Veröffentlichungen im Bereich der Sprachmodelle. Im Rahmen des Projekts wurde diese Entwicklung berücksichtigt, indem das Modell LLaMA 3.2 (veröffentlicht am 25. September 2024) integriert wurde. Durch diese Integration konnten aktuelle Fortschritte in der Sprachmodellforschung gezielt in die eigene Modellarchitektur eingebaut.

Abstimmung mit nationalen und internationalen Forschungspartnern

Während es außerhalb des Rahmens des Projekts lag, eine internationale Kooperation mit anderen Forschungspartnern direkt zu etablieren, fand ein intensiver Austausch mit Projektpartnern im Rahmen des ODELIA-Konsortiums statt. Diese Zusammenarbeit ermöglichte es, methodische Anforderungen frühzeitig zu diskutieren und sicherzustellen, dass das entwickelte System mit föderierten Lernansätzen und internationalen Datenschutzstandards kompatibel bleibt. Darüber hinaus wird eine intensivere internationale Kooperation für die Zukunft in Erwägung gezogen.

6. Veröffentlichungen

Bis zum Ende des Projekts im März 2025 lag noch keine wissenschaftliche Publikation vor. Es ist jedoch eine Veröffentlichung in einer internationalen Fachzeitschrift, wie beispielsweise European Radiology, geplant. Darüber hinaus wurde der Code für das entwickelte Modell auf GitHub veröffentlicht und steht somit bereits Forschenden zur Verfügung, die auf diese Ressource zugreifen und das Modell weiterentwickeln können.